www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2023; 12(3): 1308-1312
© 2023 TPI
www.thepharmajournal.com
Received: 13-12-2022
Accepted: 16-01-2023

Pawan Choudhary
Ph.D. Scholar, Sabarmati University (Formerly, Calorx Teachers' University), Ahmedabad, Gujarat, India

Upendra Singh
Assistant Professor, SKNCOA, Jobner, Jaipur, Rajasthan, India

US Tanwar
Assistant Professor, Govt. Agriculture College, Baharawanda, Dausa, Rajasthan, India

Vijay Singh Meena
Assistant Professor, Govt. Agriculture College, Bharatpur, Rajasthan, India

Corresponding Author:
Pawan Choudhary
Ph.D. Scholar, Sabarmati University (Formerly, Calorx Teachers' University), Ahmedabad, Gujarat, India

Support vector machines for crop disease forecasting using different approaches

Pawan Choudhary, Upendra Singh, US Tanwar and Vijay Singh Meena

Abstract
Crop diseases are major issues as it reduces the production and quality considerably and creates a major threat to food security. But due to the lack of the necessary infrastructure and knowledge, immediate identification of diseases and rectification is not being done by most of the farmers which results in heavy crop damage and hence loss to them. This paper focuses on finding out the crop disease using data mining Classification techniques based on the physical characteristics of the crop. Classification algorithms like Decision Tree, Logistic Regression, Naive Bayes, Kernel SVM, kNN and Linear SVM are deployed on plant dataset. The performances of the algorithms are analyzed based on certain metrics like Execution Time, Accuracy Score, Cohen's Kappa, Hamming Loss, Explained Variance Score, Mean Absolute Error, Mean Squared Error and Mean Squared Logarithmic Error. For analysis, Confusion matrix and Classification report are used. The Decision Tree is found to be the fastest algorithm and results in the best ATR. But the Linear SVM is found to be the best algorithm for all other metrics for the crop disease dataset. Various oversampling techniques like Synthetic Minority Oversampling Technique (SMOTE), SVM SMOTE and borderline SMOTE with sample and resample are applied to improve the performance of the Linear SVM. Linear SVM on the oversampled dataset using SVM SMOTE has a better performance for all metrics except Execution Time than that on the original dataset.

Keywords: Prediction, regression, naive Bayes, crop disease

Introduction
Due to Green revolution and advancement in technologies, it has become possible to produce enough food to the entire human society. But, factors like climatic changes (Amos *et al*, 2014) [1], crop diseases (Richard and Peter, 2005) [7], soil erosion (Factors threatening the food security, 2020), etc. are threatening the food security. Plants are being attacked by a wide range of potential microorganisms. The interaction of plant and microbe is due to microbes taking the nutrient from plant and plant defending against microbes (Mohsen *et al*, 2015) [3]. Apart from threat to food security globally, crop diseases cause a heavy loss to farmers. To avoid such loss, the farmers should identify the disease at the earlier stage and use appropriate pesticide. Hence to assist the farmers in this regard, automation of identification of crop disease is necessary.

Crop disease prediction has been done in the recent years using Machine learning methods, which showed higher accuracy than the traditional statistical methods like regression analysis. Noisy and multi-faceted data are dealt well by the machine learning methods. Support Vector Machines were used earlier for crop disease detection and Classification. Factors like soil quality, crop rotation cycle, seed quality etc. can lead to poor health and diseases in crops. Machine learning algorithms provide valuable disease classifiers by effectively taking into consideration all the possible factors, historic data as well as satellite/sensor data of fields. The study was undertaken with the following objectives of crop disease prediction.

1. To prevent crop loss by assisting farmers in identifying the disease of the crop cultivated
2. To analyze the performance of different algorithms based on various metrics on the plant dataset

To improve the performance of the best algorithm by proposing oversampling techniques to predict the disease more accurately.

Materials and Methods
Data collection and pre-processing
In this work, soybean crop disease dataset is taken from the openly available data source (Plant disease dataset, 2020).

It has 35 categorical attributes, some nominal and some ordered and 19 classes i.e., 19 types of diseases (Plant disease dataset, 2020). The collected data was pre-processed by removing unwanted data, noisy data and blank data. The data was initially coded in EXCEL and finally converted to comma delimited. CSV. Data mining Classification techniques are applied over the dataset to predict the disease in the plant.

Performance analysis of various Classification algorithms for Crop disease prediction

The soybean crop disease dataset is taken from the openly available data source. It has 35 categorical attributes, some nominal and some ordered and 19 classes i.e., 19 types of diseases. The collected data was pre-processed by removing unwanted data, noisy data and blank data. The data was initially coded in EXCEL and finally converted to comma delimited .CSV.

Experimental Simulation

The various Classification algorithms Decision Tree, Kernel SVM, kNN, Linear SVM, Logistic Regression, and Naive Bayes were executed on the processed dataset in .CSV file using PYTHON and the various metrics like Execution Time, Accuracy on Training set, Accuracy on Test set, Accuracy score, Cohen's Kappa, Hamming Loss, Explained Variance Score, Mean Absolute Error, Mean Squared Error and Mean Squared Logarithmic Error were analysed.

Results and Discussion

Table 1 shows the various metrics resulted by the various algorithms applied on the processed dataset and figures 1.(a) and 1.(b) show the comparison of the algorithms based on the maximum and minimum value metrics. As can be seen from the table 1 and figures 1.a & 1.b. The Decision Tree algorithm has cent percent Accuracy on Training Set (ATR) and minimum Execution time and for all remaining metrics, Linear SVM has produced the best results. Linear SVM stands second in the ATR. Decision Tree stands second in all metrics other than ATR and Execution time.

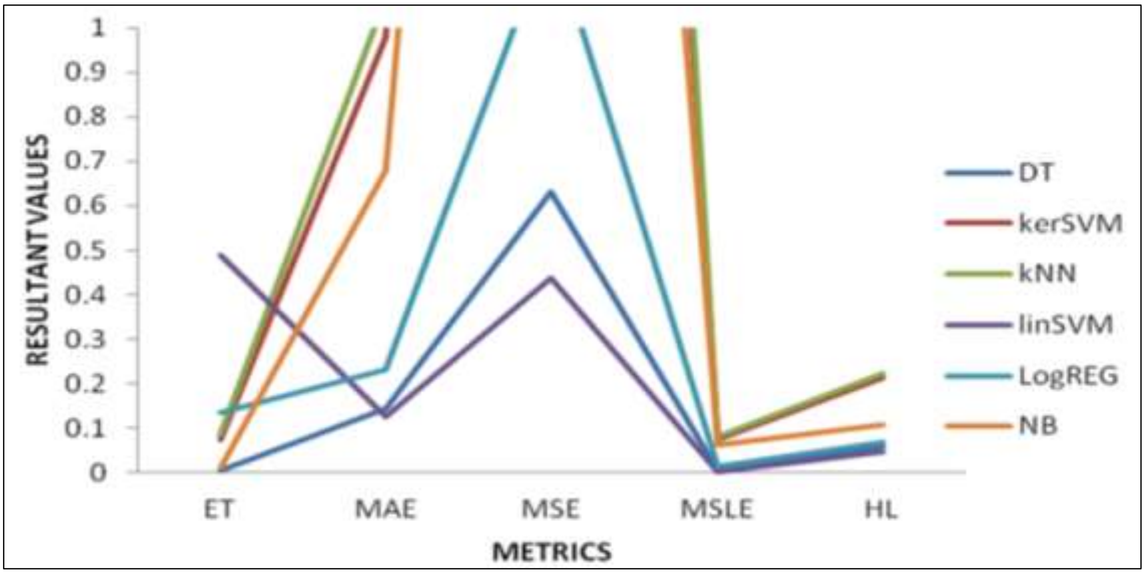

Fig 1 (a): Comparison of various algorithms based on minimum value metrics

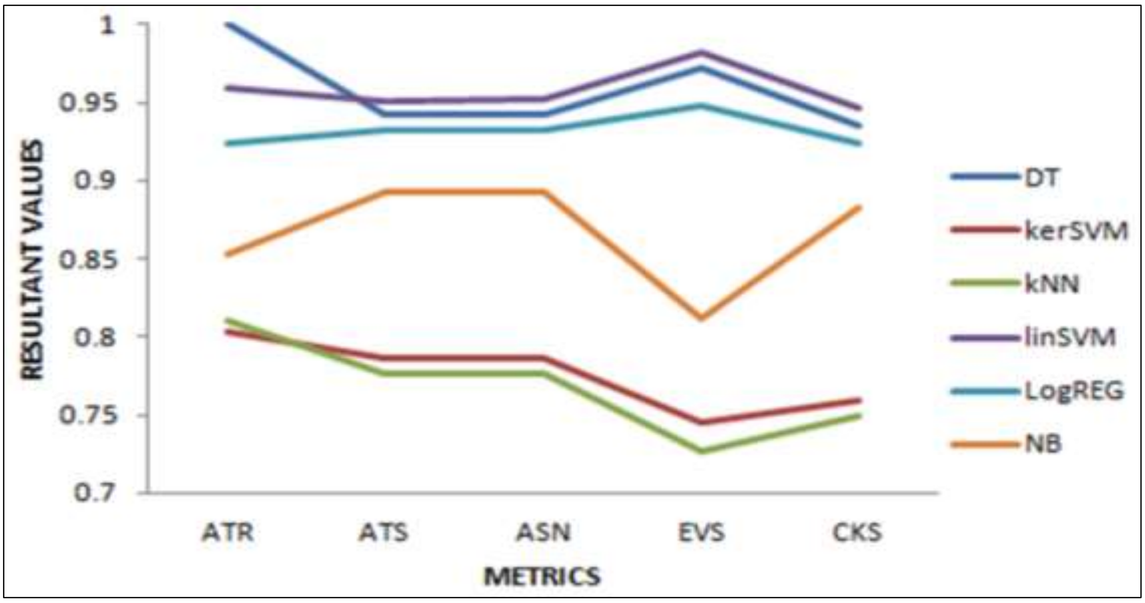

Fig 1 (b): Comparison of various algorithms based on maximum value metrics

So, for predicting crop disease, Decision Tree can be used when Execution time alone is given priority, whereas Linear SVM can be used when there is a little compromise in Execution time and ATR. As Linear SVM results in best results in most of the metrics, it is considered to be the best one for Crop disease prediction.

Table 1: Performance of the algorithms based on various metrics

Metrics	DT	KerSVM	kNN	LinSVM	LogREG	NB
ET	0.0037	0.0731	0.0854	0.4882	0.1362	0.0114
ATR	1	0.803	0.81	0.96	0.924	0.853
ATS	0.942	0.786	0.777	0.951	0.932	0.893
AS Norm-True (ASN)	0.9417	0.7864	0.7767	0.9515	0.932	0.8932
AS Norm-False (ASNF)	97	81	80	98	96	92
EVS	0.9724	0.7447	0.7261	0.9813	0.9476	0.812
MAE	0.1456	0.9806	1.068	0.1262	0.233	0.6796
MSE	0.6311	5.9903	6.3883	0.4369	1.2233	4.8155
MSLE	0.0046	0.074	0.0813	0.0029	0.0127	0.0614
CKS	0.9345	0.759	0.7494	0.9464	0.9244	0.8828
HL	0.0583	0.2136	0.2233	0.0485	0.068	0.1068

Performance analysis of Linear SVM by applying various oversampling techniques for crop disease prediction

From the previous section, it is found that Linear SVM is the best one for the crop disease dataset. In this section, its performance is improved further by using oversampling techniques. Data Imbalance occurs often in natural real-world data where one class will be having much less amount of data compared to other classes. Such imbalanced data needs to be done something so that the model results in optimal accuracy for all data classes. Some of the solutions are (i) oversampling, in which minority data is replicated, (ii) under sampling, in which majority data is reduced and (iii) hybrid of both. The net result is there should be a balance between majority and minority data (Shabrina and Joko, 2019) [8]. When a dataset is imbalanced, it may lead to inaccurate results. When the data is biased, the results will also be biased (Problems with imbalanced dataset, 2020) [8]. Data oversampling is a technique applied for imbalanced dataset to generate data similar to the underlying distribution of the real data. Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling technique that relies on the concept of nearest neighbors to create its synthetic data.

Experimental simulation

In this work, SMOTE, SVM SMOTE and borderline SMOTE are applied on the processed dataset to oversample both by sample and resample method and then the performance of the Linear SVM on these datasets are analyzed.

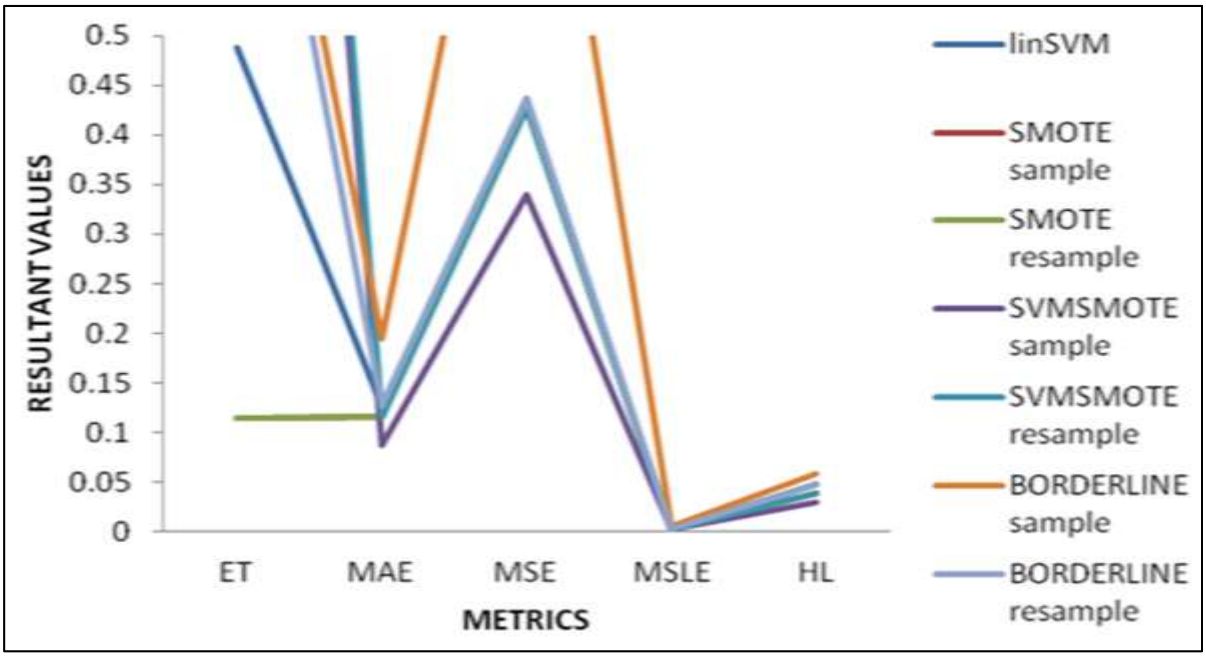

Fig 2 (a): Comparison of various algorithms based on minimum value metrics

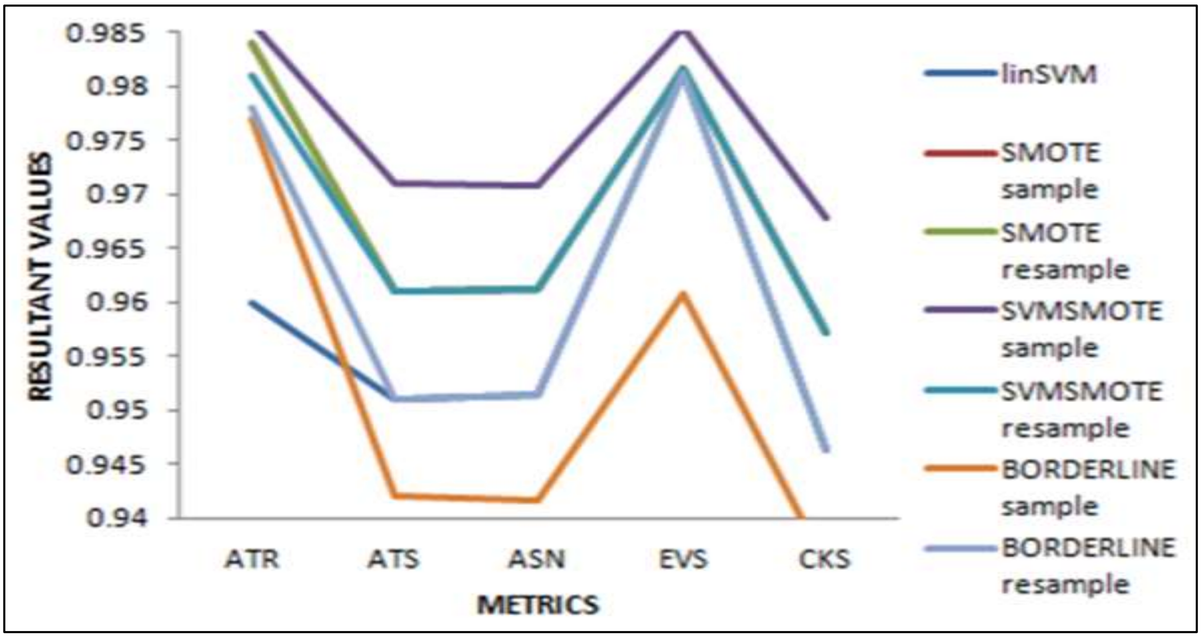

Fig 2 (b): Comparison of various algorithms based on maximum value metrics

Table 2 shows the performance of Linear SVM on the various oversampled datasets based on various metrics, figures 2.a and 2.b show the comparison of the performance of Linear SVM on the various oversampled datasets based on the maximum and minimum value metrics. As can be seen from the table 2 and figures 2.a & 2.b, the Linear SVM on SVMSMOTE with sample technique dataset has produced the best result for all metrics other than Execution Time (ET). When Execution time alone is given importance, SMOTE can be used.

Table 2: Performance of Linear SVM on the oversampled datasets based on various metrics

Metrics	Original Dataset	Smote sample	Smote resample	SVM Smote sample	SVM Smote resample	Border Line sample	Border Line Resample
ET	0.4882	0.1137	0.1137	1.8665	1.9667	0.9467	0.9042
ATR	0.9600	0.9840	0.9840	0.9860	0.9810	0.9770	0.9780
ATS	0.9510	0.9610	0.9610	0.9710	0.9610	0.9420	0.9510
ASN	0.9515	0.9612	0.9612	0.9709	0.9612	0.9417	0.9515
ASNF	98	99	99	100	99	97	98
EVS	0.9813	0.9817	0.9817	0.9854	0.9817	0.9609	0.9813
MAE	0.1262	0.1165	0.1165	0.0874	0.1165	0.1942	0.1262
MSE	0.4369	0.4272	0.4272	0.3398	0.4272	0.9126	0.4369
MSLE	0.0029	0.0028	0.0028	0.0024	0.0028	0.0061	0.0029
CKS	0.9464	0.9571	0.9571	0.9678	0.9571	0.9357	0.9464
HL	0.0485	0.0388	0.0388	0.0291	0.0388	0.0583	0.0485

Conclusion

This study has conducted a comparison between the various Classification techniques like Decision Tree, KNN, Kernel SVM, Linear SVM, Logistic regression and Naive Bayes on the crop disease dataset using PYTHON to predict the crop disease. The performances of these algorithms are analysed based on metrics like Execution Time, Accuracy on Training set, Accuracy on Test set, Accuracy score, Cohen's Kappa, Explained Variance Score, Hamming Loss, Mean Squared Error, Mean Absolute Error and Mean Squared Logarithmic Error. It is concluded that with little compromise in the Execution Time and Accuracy on the Training Set, Linear SVM is found suitable for the dataset. The Decision Tree algorithm has cent percent Accuracy on the Training Set and minimum Execution Time.

To improve the performance of the Linear SVM, the dataset is oversampled by using SMOTE, SVM SMOTE and borderline SMOTE with sample and resample, and their performances are compared. Oversampling technique SVM SMOTE with sample has a better performance for all metrics except Execution Time. When Execution Time alone is given importance, SMOTE can be used.

References

1. Amos PK Tai, Maria Val Martin and Colette L. Heald. Threat to future global food security from climate change and ozone air pollution, Nature Climate Change. 2014;4:817-821.
2. Factors threatening the food security; c2020: https://bee-inc.com/2014/10/15/threats-to-global-food-security.
3. Mohsen Mohamed Elsharkawya, Mai Nakatanib, Mitsuyoshi Nishimurab, Tatsuyuki Arakawab, Masafumi Shimizub, Mitsuro Hyakumachib. Suppression of rice blast, cabbage black leaf spot, and tomato bacterial wilt diseases by Meyerozyma guilliermondii TA-2 and the nature of protection, Acta Agriculturae Scandinavica, Section B - Soil and Plant Science; c2015. p. 1-8.
4. Plant disease dataset; c2020. https://datahub.io/machine-learning/soybean.
5. Plant disease dataset; c2020.

http://archive.ics.uci.edu/ml/datasets/Soybean+(Large).

6. Problems with imbalanced dataset; c2020. https://www.einfochips.com/blog/addressing-challenges-associated-with-imbalanced-datasets-in-machine-learning.
7. Richard N, Peter R Scott. Plant disease: a threat to global food security, Article in Annual Review of Phytopathology. 2005;43:83-116.
8. Shabrina Choirunnisa, Joko Lianto. Hybrid Method of Undersampling and Oversampling for Handling Imbalanced Data, In Proceedings of IEEE International Seminar on Research of Information Technology and Intelligent Systems; c2019. p. 276-280.